

Danmarks Tekniske Universitet

Side 1 af 11 sider

Skriftlig prøve, den 27/5-2014

Kursus navn *Introduktion til Bioinformatik*

Kursus nr. *27611*

Varighed: 4 timer

Tilladte hjælpemidler: *Alle*

"Vægtning"

Angivet ved de individuelle opgaver

27611 Eksamen Sommer 2014

Dette sæt indeholder en sammenfattende indledning og 4 opgaver (side 4-11)

Indledning – ER proteiner, KDEL motiv og KDEL receptor	side 4
Opgave 1 – Karakterisering af KDEL receptoren fra menneske og gær (25%)	side 5
Opgave 2 – Karakterisering af KDEL motivet (25%)	side 6
Opgave 3 – Fylogeni af KDEL receptoren (25%)	side 8
Opgave 4 – Forudsigelse af ER lumen proteiner (25%)	side 10

Svar til opgavesættet skal skrives enten i rå tekst (fx i JEdit) eller i et tekstbehandlingsprogram såsom Microsoft Word. **Gyldige afleveringsformater er .txt, .doc, .docx, .rtf og .pdf.** I opgave 2 og 3 skal der afleveres billeder – de kan enten indsættes i et Word dokument eller uploades separat som .png, .jpg, .gif, .eps, eller .pdf.

Svaret skal uploades på CampusNet under kursus 27611 (under "Opgaver → Sommerekksamen 2014"). **Husk at gemme seneste version af dokumentet inden du uploader svaret.** Og tjek en ekstra gang at det er det rigtige dokument, du uploader.

VIGTIGT!!!: Når du afleverer får du en afleveringskode af CampusNet. Den skal skrives på et stykke papir, som du får udleveret af eksamensvagten. Uden afleveringskode kan din besvarelse ikke blive bedømt!

Trådløst internet:

Du skal koble dig på det helt normale DTU Wireless system, ligesom til øvelserne.

Online materialer:

Lektionsplan:

http://wiki.bio.dtu.dk/teaching/index.php/27611:_Kursusplan_for_for%C3%A5r_2014

Linksamlingen til bioinformatik serverne er her:

http://wiki.bio.dtu.dk/teaching/index.php/Linksamling_for_27611

Hjælpefiler til opgave 2, 3 og 4 er tilgængelige via links i de enkelte opgaver, men kan også ses her:

<http://www.cbs.dtu.dk/dtucourse/27611spring2014/a/>

BEMÆRK:

- I er ikke begrænset til kun de links der findes her – det er tilladt at søge information

andetsteds.

- Det er **IKKE** tilladt at kommunikere med andre over nettet under eksamen.
- Der vil blive taget stikprøver af netværkstrafikken for at sikre dette.

Hvad gør man hvis en web-server ikke virker:

- 1) Verificer at input-data er i korrekt format. Forkert inputdata er i næsten alle tilfælde årsagen til problemet.
- 2) Prøv evt. at finde en alternativ server med samme funktion (Søg i Google).
- 3) Rapporter fejlen til eksamensvagten - den kursusansvarlige vil så blive tilkaldt.

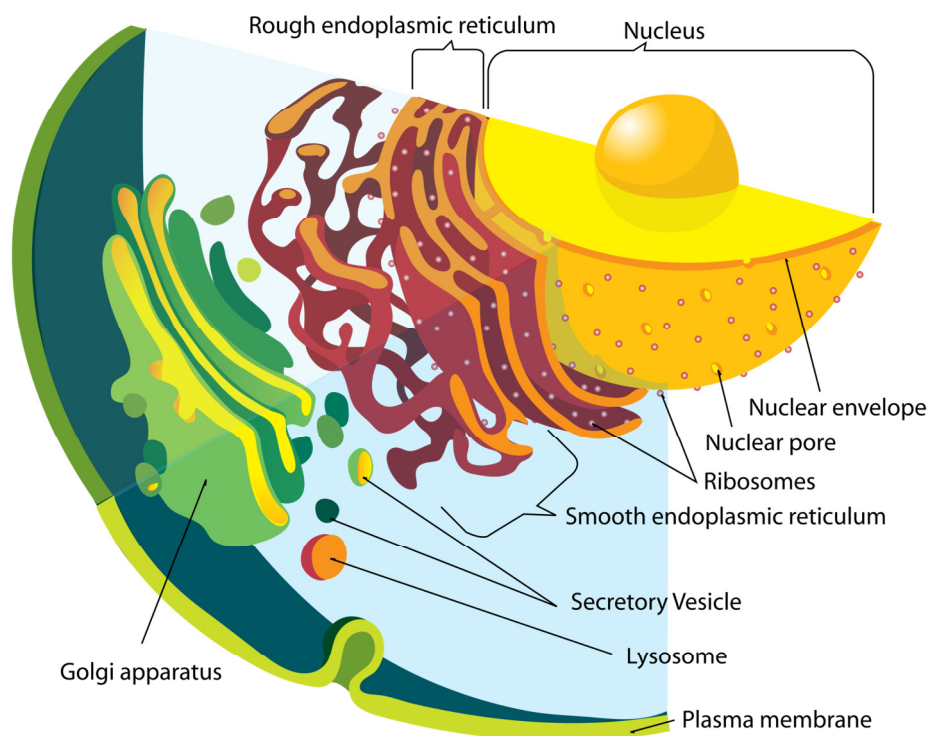
HUSK altid:

Don't panic

Held og lykke med eksamen.
Bent, Rasmus og Henrik

Indledning - ER proteiner, KDEL motiv og KDEL receptor

Eksamenssættet har et tema i år: proteiner i det endoplasmatiske reticulum (ER) og den måde, de sorteres på.



Eukaryote celler har mange forskellige organeller, hver med en karakteristisk sammensætning af proteiner. Proteiner med et signalpeptid indgår i den sekretoriske pathway, og de bliver i første omgang sendt gennem membranen til det endoplasmatiske reticulum (ER). Herfra transporteres de som hovedregel videre gennem Golgi-apparatet til celleoverfladen og cellens omgivelser. De proteiner, der hører hjemme i ER, skal aktivt holdes tilbage i ER for at forhindre, at de bliver sendt videre.

De fleste af de proteiner, der findes opløst i hulrummet (lumen) af ER (altså ikke i ER membranen), har et C-terminalt signal, der fortæller, at de skal holdes tilbage. Dette signal kaldes KDEL motivet, fordi det har konsensus-sekvensen K-D-E-L (dvs. de sidste fire aminosyrer i proteinet er oftest Lys-Asp-Glu-Leu). Det genkendes af et protein, der går under navnet "KDEL receptor" eller (mere korrekt) "ER lumen protein retaining receptor".

Opgave 1 - Karakterisering af KDEL receptoren fra menneske og gær (25%)

Spørgsmål a: Hvor mange proteiner i UniProt hedder "ER lumen protein retaining receptor"? Hvor mange af disse er fra Swiss-Prot, og hvor mange fra TrEMBL?

Husk: Når du besvarer UniProt spørgsmål, skal du angive den søgestreng, du brugte for at nå frem til svaret.

Spørgsmål b: Find "ER lumen protein retaining receptor" fra menneske (*Homo sapiens*) i Swiss-Prot. Hvor mange er der, og hvilken søgestreng brugte du?

Spørgsmål c: Se nærmere på det hit fra spørgsmål b, der hedder "ER lumen protein retaining receptor 1". Hvad er dets accession-kode, UniProt ID (Entry name) og gen-navn(e)?

Spørgsmål d: Er dette protein et transmembranprotein? Hvor i UniProt kan du se det? Hvis ja, hvor mange gange går det igennem membranen (hvor mange transmembran-segmenter har det)? Er positionen af transmembran-segmenterne eksperimentelt bestemt?

Spørgsmål e: Find nu "ER lumen protein retaining receptor" fra bagegær (*Saccharomyces cerevisiae*) i Swiss-Prot. *Bemærk at når du søger på en organisme, hvorfra der findes mange undertyper (stammer/isolater), skal du passe på ikke at komme til at begrænse søgningen til en specifik undertype.* Hvad er dens accession-kode, UniProt ID (Entry name) og gen-navn? Hvilken søgestreng skal man bruge for at finde lige præcis ét hit?

Spørgsmål f: Se på hvad der står i UniProt om funktionen af menneske-proteinet fra spørgsmål c og gær-proteinet fra spørgsmål e. Er der forskel på, hvilket motiv de genkender? Hvis ja, hvilken forskel?

Spørgsmål g: Lav et globalt parvis alignment af menneske-proteinet fra spørgsmål c og gær-proteinet fra spørgsmål e. Indsæt selve alignmentet i din besvarelse. Hvad er % identitet og % similaritet? Hvor mange gaps er der?

Spørgsmål h: Brug nu BLAST til at finde ud af, om der er et *signifikant* match mellem menneske-proteinet fra spørgsmål c og gær-proteinet fra spørgsmål e: Hvis du bruger menneske-proteinet til at søge i bagegær, finder du så gær-proteinet fra spørgsmål e som første hit? Hvad er E-værdien? Hvad er % identitet og % similaritet? Hvor mange gaps er der? *Husk at dokumentere, hvordan du har gjort, og begrund de valg, du foretager.*

Opgave 2 - Karakterisering af KDEL motivet (25%)

Spørgsmål a: Hvor mange proteiner i UniProt findes i ER ("endoplasmic reticulum")? (*Tip:* Husk at du i UniProt kan søge på en specifik "subcellular location").

Husk fortsat: Når du besvarer UniProt spørgsmål, skal du altid angive den søgestreng, du brugte for at nå frem til svaret.

Spørgsmål b: Hvor mange af proteinerne i ER (fra spørgsmål a) har eksperimentel evidens for at de findes i ER?

Spørgsmål c: Hvor mange af proteinerne i ER (fra spørgsmål a) findes i hulrummet (*lumen*) af ER?

Spørgsmål d: Hvor mange af proteinerne i ER (fra spørgsmål a) har et annoteret signalpeptid?

Spørgsmål e: Hvor mange af proteinerne i ER (fra spørgsmål a) er hele sekvenser (dvs. ikke fragmenter)?

Spørgsmål f: Lav nu en søgning efter hele proteiner fra ER lumen med signalpeptider (dvs. proteiner der opfylder kriterierne fra spørgsmål c, d og e). Hvor mange får du?

Spørgsmål g: Som vi så i opgave 1f, er der forskel på KDEL motivet mellem menneske og bagegær. Menneske og bagegær tilhører forskellige riger (*Kingdoms*). Vores hypotese er derfor, at motivet afhænger af riget. Find de videnskabelige betegnelser på de to *Kingdoms*, som hhv. menneske og bagegær tilhører (brug f.eks. NCBI Taxonomy).

Spørgsmål h: Du skal nu opdele din søgning fra spørgsmål f, så du får et datasæt fra hvert af de to riger fra spørgsmål g. Hvor mange får du fra hvert af de to riger? (*Tip:* brug de videnskabelige betegnelser fra spørgsmål g). *Hvis du ikke kunne løse spørgsmål f, så tag udgangspunkt i søgningen fra spørgsmål a i stedet..*

opgaven fortsætter på næste side...

Spørgsmål i: Du skal nu undersøge mønstret af aminosyrer i slutningen af sekvensen for proteiner fra hvert af de to riger. Vi har hjulpet lidt og lavet to filer, der indeholder *de 15 sidste aminosyrer fra hvert protein* i de to søgninger, du skulle lave i spørgsmål h.

Her er den, der indeholder menneske:

<http://www.cbs.dtu.dk/dtu/course/27611spring2014/a/cde-dyr.w15.pep>

og her er den, der indeholder gær:

<http://www.cbs.dtu.dk/dtu/course/27611spring2014/a/cde-svampe.w15.pep>

Brug EasyPred til at lave to logoer ud fra de to filer. Sæt Weight on prior (β) til 10, og brug "Cluster at 62% identity".

Aflever de to logoer som billeder (direkte i dit svar, eller som filer ved siden af).

Betragt de to logoer. Hvilke to positioner indeholder mest information? Kan man se konsensus-sekvenserne for KDEL-motiverne i dem? Beskriv forskellene mellem de to riger.

Opgave 3 - Fylogeni af KDEL receptoren (25%)

Spørgsmål a: På baggrund af den søgning efter "ER lumen protein retaining receptor", der blev foretaget i opgave 1a, er det muligt at bygge et datasæt, som kan bruges til at undersøge slægtskabet mellem de organismer, hvorfra vi kan finde KDEL receptorer. Vi skal bruge alle proteiner, der hedder "ER lumen protein retaining receptor" i *Swiss-Prot*. Du kan selv downloade dem i FASTA format fra UniProt via din søgning i opgave 1a, eller du kan bruge nedenstående FASTA fil:

<http://www.cbs.dtu.dk/dtucourse/27611spring2014/a/KDEL-receptor.fasta>

For at gøre analysen lidt lettere, er din første opgave at rydde lidt op i datasættet:

1. Fjern alle sekvenser, som er beskrevet som "putative".
2. Tilføj til hvert FASTA navn det videnskabelige navn på den overordnede gruppe, som organismen hører til. For dyr, planter og svampe skal du bruge rige ("*Kingdom*") — her må du slå op i NCBI Taxonomy, hvad de videnskabelige navne på rigerne er. For de organismer, som ikke hører til noget *Kingdom*, skal du bruge de navne på overordnede grupper, der fremgår af oversigten på næste side. Sæt teksten sammen som vist i eksemplet nedenfor — der må IKKE være mellemrum i FASTA navnet.
3. For at undgå, at navnene bliver for lange, skal du fjerne "sp|" og Accession kode fra FASTA navnet (se nedenstående eksempel).
4. Fjern desuden alle kommentarer fra FASTA filen (det, der kommer efter FASTA navnet).

Eksempel:

```
>sp|Q86JE5|ERD2_DICDI ER lumen protein retaining receptor  
OS=Dictyostelium discoideum GN=kdelr PE=3 SV=1
```

Ændres til:

```
>ERD2_DICDI_Amoebozoa
```

Indsæt din tilrettede FASTA fil i dit svar eller vedlæg den til din besvarelse som en separat fil.

Spørgsmål b: Lav et multiple alignment og derefter et fylogenetisk træ af sekvenserne i din FASTA fil. Rodfæst (*reroot*) træet ved at bruge slægten *Plasmodium* som *outgroup*. Gør teksten i træet mere læselig ved at forøge fontstørrelsen (*Tip*: dette gøres under punktet "Tip Labels" i menuen i venstre side af FigTree). Beskriv hvordan du har gjort og aflever et billede af træet.

Spørgsmål c: Betragt den overordnede fylogeni i træet. Er rigerne "dyr", "svampe" og "planter" repræsenteret som systematiske grupper (under-træer), som de bør være, eller er der uoverensstemmelser/fejl her (i så fald hvilke)? Er der andre fejl i den overordnede fylogeni (altså ikke inden for rigerne)? Brug evt. NCBI Taxonomy, hvis du er i tvivl om den korrekte systematik.

opgaven fortsætter på næste side...

Spørgsmål d: Betragt nu fylogenen inden for dyrene i træet. Læg mærke til, at nogle dyr er repræsenteret ved to eller tre sekvenser. *Bemærk:* I de tilfælde hvor en organisme har flere versioner af proteinet, vil det fremgå af det korte navn, f.eks:

ERD21_RAT = ER lumen protein retaining receptor 1

ERD22_RAT = ER lumen protein retaining receptor 2

Hvilket af følgende udsagn passer bedst med hvad du kan se i træet?

1. Opdelingen i ER lumen protein retaining receptor 1, 2 og 3 er sket meget tidligt i dyrenes udvikling, før de forskellige rækker/phyla — såsom hvirveldyr (*Vertebrata*), leddyr (*Arthropoda*) og rundorme (*Nematoda*) — spaltede ud.
2. Opdelingen i ER lumen protein retaining receptor 1, 2 og 3 er sket tidligt i hvirveldyrenes (*Vertebrata*) udvikling, før de forskellige klasser — såsom benfisk (*Actinopteri*), padder (*Amphibia*) og pattedyr (*Mammalia*) — spaltede ud.
3. Opdelingen i ER lumen protein retaining receptor 1, 2 og 3 er sket forholdsvis tidligt i pattedyrenes (*Mammalia*) udvikling, før de forskellige ordner — såsom primater (*Primates*) og gnavere (*Rodentia*) — spaltede ud.
4. Opdelingen i ER lumen protein retaining receptor 1, 2 og 3 er sket sent i dyrenes udvikling og er foregået separat i hver orden.

Begrund dit svar.

Oversigt over de arter der indgår i vores datasæt::

<u>Kort navn</u>	<u>artsnavn</u>	<u>overordnet gruppe</u>
ARATH	<i>Arabidopsis thaliana</i>	
BOVIN	<i>Bos taurus</i>	
CAEBR	<i>Caenorhabditis briggsae</i>	
CAEEL	<i>Caenorhabditis elegans</i>	
CHICK	<i>Gallus gallus</i>	
DANRE	<i>Danio rerio</i>	
DICDI	<i>Dictyostelium discoideum</i>	Amoebozoa
DROME	<i>Drosophila melanogaster</i>	
ENTHI	<i>Entamoeba histolytica</i>	Amoebozoa
HUMAN	<i>Homo sapiens</i>	
KLULA	<i>Kluyveromyces lactis</i>	
MOUSE	<i>Mus musculus</i>	
PETHY	<i>Petunia hybrida</i>	
PLAF7	<i>Plasmodium falciparum</i> (isolate 3D7)	Apicomplexa
PLAFA	<i>Plasmodium falciparum</i>	Apicomplexa
RAT	<i>Rattus norvegicus</i>	
SCHPO	<i>Schizosaccharomyces pombe</i>	
XENLA	<i>Xenopus laevis</i>	
XENTR	<i>Xenopus tropicalis</i>	
YEAST	<i>Saccharomyces cerevisiae</i>	

Opgave 4 - Forudsigelse af ER lumen proteiner (25%)

I denne opgave skal vi se, om de datasæt, vi lavede i opgave 2, kan bruges til at forudsige, om et protein hører hjemme i ER lumen eller ej. Til det formål skal vi bruge EasyPred til at lave en vægtmatrix, og det vil være nødvendigt at dele vores data op i træning og test.

Spørgsmål a: Download datasættet fra opgave 2i af de sidste 15 aminosyrer af ER lumen proteiner fra dyr:

<http://www.cbs.dtu.dk/dtucourse/27611spring2014/a/cde-dyr.w15.pep>

til din egen computer.

Brug så jEdit (eller en anden tilsvarende plain text editor) til at dele filen op i to dele, sådan at den første del (*træningssættet*) indeholder de første 200 linjer, og den anden del (*det positive testsæt*) indeholder resten. Gem hver del som en ny fil med et passende navn. Hvor mange linjer indeholder det positive testsæt? Som en kontrol af, at du har gjort det rigtigt, skal du indsætte de første 10 linjer af det positive testsæt i din besvarelse.

Spørgsmål b: Som en kontrol af, at du forstår, hvordan værdierne i en vægtmatrix udregnes, skal du manuelt beregne vægtene for aminosyrerne D og E i den tredjesidste position (position 13). Du skal dog ikke bruge hele træningssættet, men kun de første 10 eksempler. De er for nemheds skyld sat ind her:

```
GRENSFKGFHRRNEF 1
RHEHYFNQFHRRNEL 1
HRHNHFSRFHRQDEL 1
DETGDAAPVEEHDEL 1
EEEEEDAAAGQAKDEL 1
DEEEEKEEEEKGHDEL 1
DDEEEEKEEEEKGHDEL 1
EDEEEDVPGQAKDEL 1
EDEEDTTPGQTKDEL 1
DESKQDKDQSEHDEL 1
```

Sæt Weight on prior (β) til 10, og brug ikke clustering. Husk at angive mellemregninger i din besvarelse (f -, g -, og p -værdier) og ikke blot det færdige resultat (w -værdier).

Tip: Du får brug for tabellerne i handoutet om beregning af *pseudocounts*, det ligger her:

<http://www.cbs.dtu.dk/courses/27625.algo/presentations/PSSM/Estimationofpseudocounts.pdf>

Spørgsmål c: Vi skal også bruge et negativt testsæt til at måle, om vores vægtmatrix kan forudsige noget. Til det formål skal vi samle proteiner, der bevæger sig gennem ER, men ikke bliver holdt tilbage af KDEL receptoren. Lav en UniProt søgning efter proteiner fra dyr, som har signalpeptider men ikke hører hjemme i ER ("endoplasmic reticulum"). Undgå, ligesom før, at tage fragmenter med. Tænk ikke på hvorvidt der er eksperimentel evidens. Hvor mange får du? Husk (som altid) at angive søgestrengen i din besvarelse.

opgaven fortsætter på næste side...

Spørgsmål d: Det datasæt, du lige har samlet, er lige lovlig stort til at bruge som negativt testsæt. Vi har derfor lavet en udgave, som kun indeholder hver 100^{te} sekvens af det, og så har vi, ligesom før, taget de sidste 15 aminosyrer af hver sekvens. Det ligger her:

<http://www.cbs.dtu.dk/dtucourse/27611spring2014/a/negative-dyr.w15.test0.pep>

Læg mærke til, at vi har tilføjet et nul i hver linje for at vise, at dette er negative data, i modsætning til de positive data, som alle har et 1-tal.

Brug nu jEdit (eller tilsvarende) til at kombinere det positive testsæt fra spørgsmål a og dette negative testsæt til et samlet testsæt (også kaldet evalueringssæt), som du gemmer som en ny fil med et passende navn. Hvor mange linjer har dit samlede testsæt?

Spørgsmål e: Indsæt nu træningssættet og testsættet i EasyPred. Sæt Weight on prior (β) til 10, og brug "Cluster at 62% identity". Se på resultaterne af forudsigelsen. Hvad er Pearson korrelationskoefficienten, og hvad er Aroc værdien? Hvad skulle de være, hvis vores metode fungerede perfekt? Hvad ville de være, hvis vores metode bare gættede tilfældigt?